NBDI Data Analyst Trainee Case Assignment 2026

Case Information:

Time limit: approximately 1 hour (maximum 90 minutes)

Submission deadline: Thursday at 20:00 (Finnish time)

Submission format: one script (solution.R or solution.sql) and one CSV output file (education_diversity_by_company.csv)

Submit by email to aaro.angerpuro@impaktly.com

The following case assignment is part of the recruitment process for the Nordic Business Diversity Index (NBDI) 2026. The purpose of this short exercise is to understand how you approach data cleaning and analysis using either R or SQL. The task is designed to take around one hour in total and should not exceed ninety minutes. The focus is on clear logic, reproducibility, and a structured way of thinking about data.

You will work with the file “Stockholm Large – NBDI2025.xlsx.” The file includes information about companies and their leadership teams, including each individual’s position, gender, nationality, and educational background. The aim of the exercise is to examine how diverse executive management teams are in terms of education.

Use the sheet named “Data”. The relevant columns are Company, Board of Directors or Executive Management, Position, Educational Background, Gender, Nationality, and Year of Birth.

Your task is to clean the data, focus only on executive management, and produce one summary table that shows educational diversity by company. Begin by standardising the variable “Board of Directors or Executive Management” so that it only contains the values “Board” or “Executive.” Keep only the rows that belong to “Executive.” Standardise gender into “Male,” “Female,” or “Other/Unknown.” Normalise the educational background into one or more of the following categories: Business, Law, Engineering, Sciences, Humanities, Medicine, Arts, or Other: “specific degree.” If a person’s education is not listed, write “N/A.” Do not guess or invent degrees. These cleaning steps must be done with code, not manual editing or excel functions.

Once the data is cleaned, calculate a summary of educational diversity for each company’s executive team. For each company, show the number of executives, the number of unique education categories represented, the share of the most common education field, and a simple diversity score.

Continues…

For each company, calculate a simple score that shows how mixed the executives' educational backgrounds are. To do this, first calculate how many executives fall into each education category and then how large each group's share is. For example, if a company has five executives and three of them studied business, the share for business is 0.6. Next, square each of those shares (for example, 0.6 × 0.6 = 0.36), add them together, and subtract that total from one. The resulting number is the diversity score. If everyone studied the same thing, the score will be close to zero. If the team includes people with many different educational backgrounds, the score will be higher, approaching one. This method automatically adjusts for the number of executives or education types a company has. If some executives have more than one degree, you can count each degree separately, as long as you base the shares on the total number of degrees rather than the number of people. The goal is simply to measure how varied the education mix is within each executive team.

For each company, calculate the diversity score using this formula:

$$\text{Diversity Score} = 1 - \Sigma(p_i^2)$$

where p_i is the share of executives (or degrees) that belong to education category *i*.

In plain terms:

- Find how many executives (or total degrees) fall into each education category.
- Divide each count by the total number of executives (or total number of degrees) to get the shares.
- Square each share, add them together, and subtract that total from one.

If everyone studied the same field, the score will be **0** (no diversity).
If the team has many different educational backgrounds with similar proportions, the score will be closer to **1** (high diversity).

Continues…

Produce one output table named “education_diversity_by_company.csv.” Your script should read the Excel file, clean the data, and generate this single output. Submit your script as “solution.R” or “solution.sql,” depending on which tool you choose. You may optionally include a short text note of up to one hundred words explaining your approach or any assumptions you made.

You can complete the task in either R or SQL. In R, you may use the tidyverse and readxl packages. In SQL, you may use any dialect such as PostgreSQL, SQLite, or MySQL. CSV import from Excel is acceptable. The important thing is that your work is fully reproducible, and your code can be run from start to finish without manual intervention. Aim for clarity and simplicity, not perfection.

Submissions will be evaluated on the logic of your data cleaning, the correctness of your calculations, the clarity of your output, and whether the code runs cleanly. The goal is not to test advanced syntax, but to see how you think about structure and data.

Good luck with the case.

Aaro Angerpuro
Impaktly